BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Natural Language Processing (AIMLC ZC)

Dr Naveen Kumar Laskari
laskari@wilp.bits-pilani.ac.in

These slides are prepared by the instructor, with grateful acknowledgement of James Allen and many others who made their course materials freely available online.

Session 1
(26th July 2026)

Introduction to NLP

Agenda for the session

- Objective of course
- What will we learn in this course?
- Text books and Reference books
- Evaluation Plan
- What is Natural Language Processing?
- Application areas of Natural Language Processing
- Introduction to Natural Language Processing

Objective of course

- To Identify and recall fundamental concepts and techniques in Natural Language Processing (NLP).
- To Explain the computational properties of natural languages and articulate the algorithms commonly utilized for processing linguistic information.
- To Apply basic mathematical models and methods in NLP applications to solve problems and perform tasks.
- To Examine research and development efforts in Natural Language Processing to discern trends, challenges, and opportunities.

What you will discuss in this course

- Vector Semantics
- Word Embedding
- Language Modelling
  - N-gram language modeling
  - Neural Language Models
- Introduction to LLM and Prompt Engineering
- Part-of-Speech Tagging
- Parsing
- Encoder-Decoder Models, Attention and Contextual Embedding's, BERT
- Word sense disambiguation
- Semantic web ontology and knowledge Graphs
- Retrieval Augmented Generation (RAG)
- State of the art NLP applications

Text books and Reference books

T1	Jurafsky and Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, McGraw Hill
T2	Manning and Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA

R1	Allen James, Natural Language Understanding
R2	Neural Machine Translation by Philipp Koehn
R3	Semantic Web Primer (Information Systems) By Antoniou, Grigoris; Van Harmelen, Frank

Evaluation Plan

Name	Weight
Quiz (best 2 out of 3)	10%
Assignment 1 and 2	20%
Mid-term Exam	30%
End Semester Exam	40%

What is Natural Language Processing?

- Natural Language Processing (NLP) is a specialized branch of Artificial Intelligence (AI) that gives computers the ability to read, interpret, understand, and generate human language.
- It sits at the intersection of computer science, artificial intelligence, and computational linguistics

What is Natural Language Processing?

- Natural Language Processing
  - Process information contained in natural language text.
  - Also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)
  - Analyze, understand and generate human languages just like humans do.
  - Applying computational techniques to language domain..
  - To explain linguistic theories, to use the theories to build systems that can be of social use..
  - Started off as a branch of Artificial Intelligence..
  - Borrows from Linguistics, Psycholinguistics, Cognitive Science & Statistics.
  - Make computers learn our language rather than we learn theirs.

Why Study NLP?

- A hallmark of human intelligence.
- Text is the largest repository of human knowledge and is growing quickly.
  - emails, news articles, web pages, IM, scientific articles, insurance claims, customer complaint letters, transcripts of phone calls, technical documents, government documents, patent portfolios, court decisions, contracts, ……

➢ **The Natural Language Processing (NLP) Market size to grow from USD 15.7 billion in 2022 to USD 49.4 billion by 2027**, at a Compound Annual Growth Rate (CAGR) of 25.7% during the forecast period.

Why are language technologies needed?

- Many companies make a lot of money if they could use computer programmes that understood text or speech.
  - answering the phone, and replying to a question
  - understanding the text on a Web page to decide who it might be of interest to
  - translating a daily newspaper from Japanese to English
  - understanding text in journals / books and building an expert systems based on that understanding

From Dreams to reality

Yesterday's Dream	Today's Reality
Computers that can talk like humans	ChatGPT, Gemini, Claude
Instant language translation	Real-time multilingual communication
Read thousands of articles and summarize them	AI-generated summaries in seconds
Answer questions like a human expert	Conversational AI assistants
Write emails, reports, and essays	AI-powered content generation
Generate and explain code	AI coding assistants
Personalized tutoring	AI learning companions

Why Do We Still Need NLP?

AI is powerful... but not perfect.

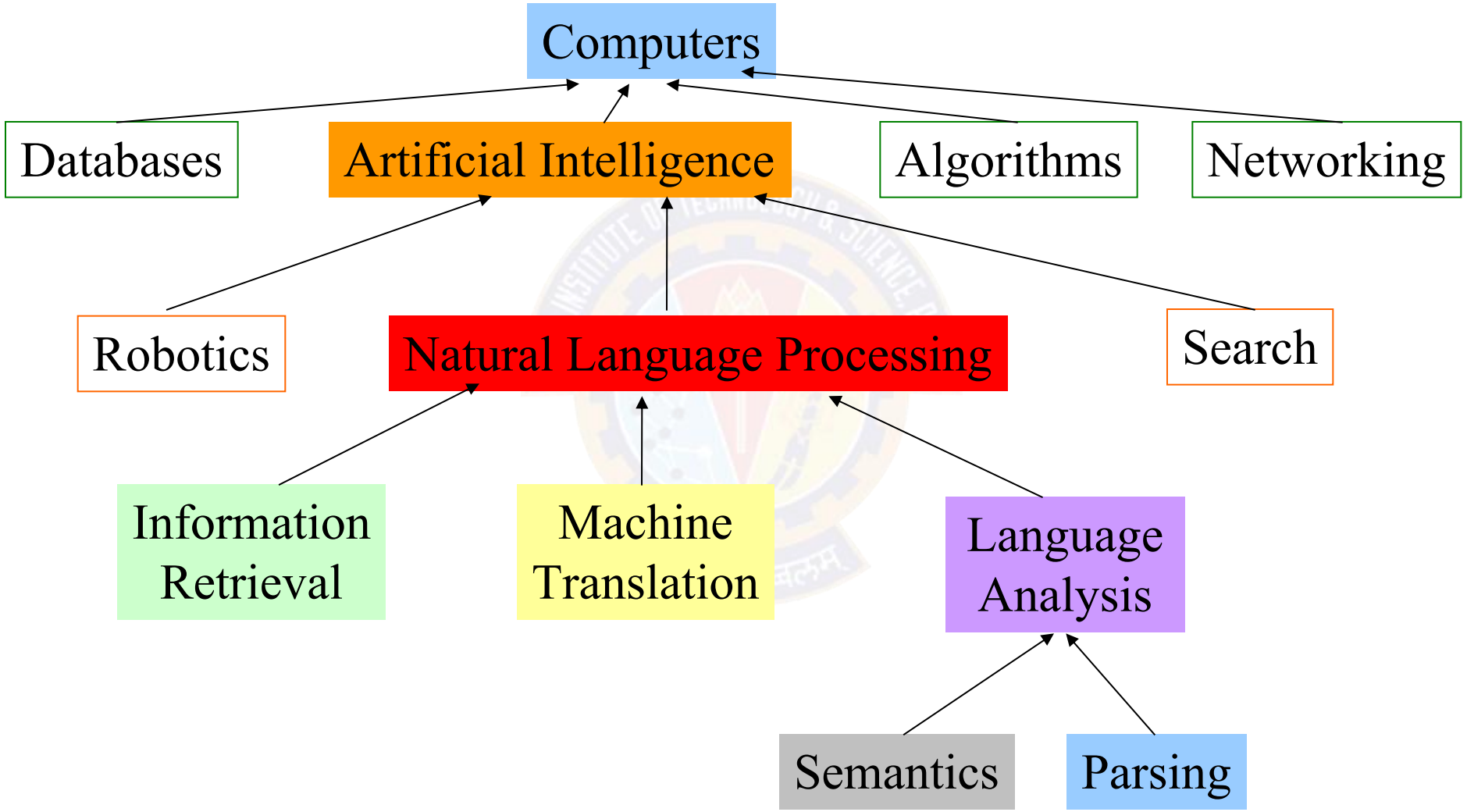
AI can

- Understand natural language
- Summarize large documents
- Translate between languages
- Answer questions
- Generate text and code
- Converse naturally

Yet AI still struggles with

- Ambiguity and sarcasm
- Context over long conversations
- Hallucinations (confident but incorrect answers)
- Bias and fairness
- Domain-specific knowledge
- Explainability and trustworthiness

Where does it fit in the CS taxonomy?



Evolution of NLP

- **Natural Language Processing (NLP)** evolved from strict rules of grammar to statistical probabilities, and finally to modern self-learning transformer models. This evolution occurred over four major eras.
  - The Symbolic Era (1950s–1980s)
  - The Statistical Era (1990s–2000s)
  - The Neural Network Era (2010s–2015)
  - The Transformer and LLM Era (2017–Present)

The Symbolic Era (1950s–1980s)

The foundations of NLP were built on hardcoded linguistic rules and symbolic logic.

- **1950:** Alan Turing published his famous paper "Computing Machinery and Intelligence," introducing the **Turing Test** as a benchmark for machine intelligence
- **1954:** The Georgetown-IBM experiment successfully translated 60 Russian sentences into English automatically. This sparked massive, overoptimistic funding for machine translation
- **1966:** The **ALPAC report** concluded that machine translation was more expensive and less accurate than human translation. This caused funding to collapse, triggering the first "AI Winter.“
- **1966:** Joseph Weizenbaum created **ELIZA**, the world’s first chatbot. It simulated a psychotherapist by matching user keywords to pre-written script responses

The Statistical Era (1990s–2000s)

NLP shifted from telling computers how language worked to letting computers calculate language patterns using data.

- **Data-Driven Approaches:** Instead of linguists writing thousands of grammar rules, algorithms analyzed massive collections of text called text corpora
- **Core Algorithms:** Researchers introduced **Hidden Markov Models (HMM)** and **Statistical Machine Translation (SMT)**
- **The Shift:** Systems calculated the mathematical probability that a specific word would follow another, vastly improving translation and speech recognition.

The Neural Network Era (2010s–2015)

The introduction of Deep Learning allowed computers to learn complex features of language without manual feature engineering

- **Word Embeddings:** In 2013, Google researchers introduced Word2Vec. This technique mapped words into numerical vectors, allowing machines to understand that "king" and "queen" are mathematically related
- **Sequence Models: Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM)** networks became the standard. They allowed models to process text sequentially and remember previous words in a sentence

The Transformer and LLM Era (2017–Present)

Modern NLP relies on models that process entire blocks of text simultaneously, capturing deep context

- **2017:** Google researchers published the groundbreaking paper "*Attention Is All You Need,*" introducing the **Transformer architecture**. This eliminated the need for sequential processing, letting models look at all words at once.
- **2018:** Google released **BERT**, which read text bidirectionally (left-to-right and right-to-left) to capture deep context
- **2020–2021:** OpenAI launched the **GPT** series (Generative Pre-trained Transformer), shifting the NLP paradigm toward scaling up parameters and massive datasets. This gave rise to **Large Language Models (LLMs)** capable of advanced reasoning, coding, and creative text generation

LLM and Multimodal models

- 2020: GPT-3 shows large language models can learn new tasks from just a few examples (“few-shot learning”)
- Nov 2022: ChatGPT launches — conversational LLMs reach mainstream adoption
- Mar 2023: GPT-4 adds multimodal (text + image) input and stronger reasoning
- 2023: Retrieval-Augmented Generation (RAG) becomes the standard way to ground LLMs in external, up-to-date knowledge
- 2023–2024: Open-weight models (e.g. Llama, Mistral) broaden access to LLM technology
- Sept 2024: Dedicated “reasoning” models (e.g. OpenAI o1) introduce explicit step-by-step reasoning at inference time
- 2024–2026: Agentic AI — LLMs that plan, call tools, and act autonomously — becomes a major research and industry focus
- 2025–2026: Frontier labs continue rapid iteration, converging reasoning and multimodal capability into single flagship models

The study of language

Discipline	Typical Problems	Tools
Linguists	How do words form phrases and sentences? What constrains the possible meanings for a sentence?	Intuitions about well-formedness and meaning; mathematical models of structure and meaning
Psycholinguists	How do people identify the structure of sentences? How are word and text meanings identified?	Experimental techniques based on measuring human performance; statistical analysis of observations
Philosophers	What is meaning, and how do words and sentences acquire it? How do words identify objects in the world?	Natural language argumentation using intuition about counter-examples; mathematical models (for example, logic and model theory)
Computational Linguists	How is the structure of sentences identified? How can knowledge and reasoning be modeled? How can language be used to accomplish specific tasks?	Algorithms, data structures; formal models of representation and reasoning; AI techniques (search and representation methods)

Beyond Words – What Makes AI Intelligent?

Dave: Open the pod bay doors, HAL.
HAL: I am sorry, Dave. I am afraid I can't do that.
Dave: What's the problem.
HAL: I think you know what the problem is just as well as I do.
Dave: I don't know what you're talking about.
HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.

What capabilities does HAL demonstrate?

- Natural language understanding – understands and responds appropriately.
- Reasoning – infers hidden intentions from available information.
- Goal-oriented decision making – acts according to its objectives.
- Theory of Mind – models human beliefs, intentions, and future actions.
- Social intelligence – participates in a coherent conversation.
- Context awareness – uses previous interactions to interpret meaning.

True NLP is not just processing words—it is understanding meaning, context, intentions, and human behavior.

From Language Processing to Language Understanding

HAL's Advanced Human-Like Abilities

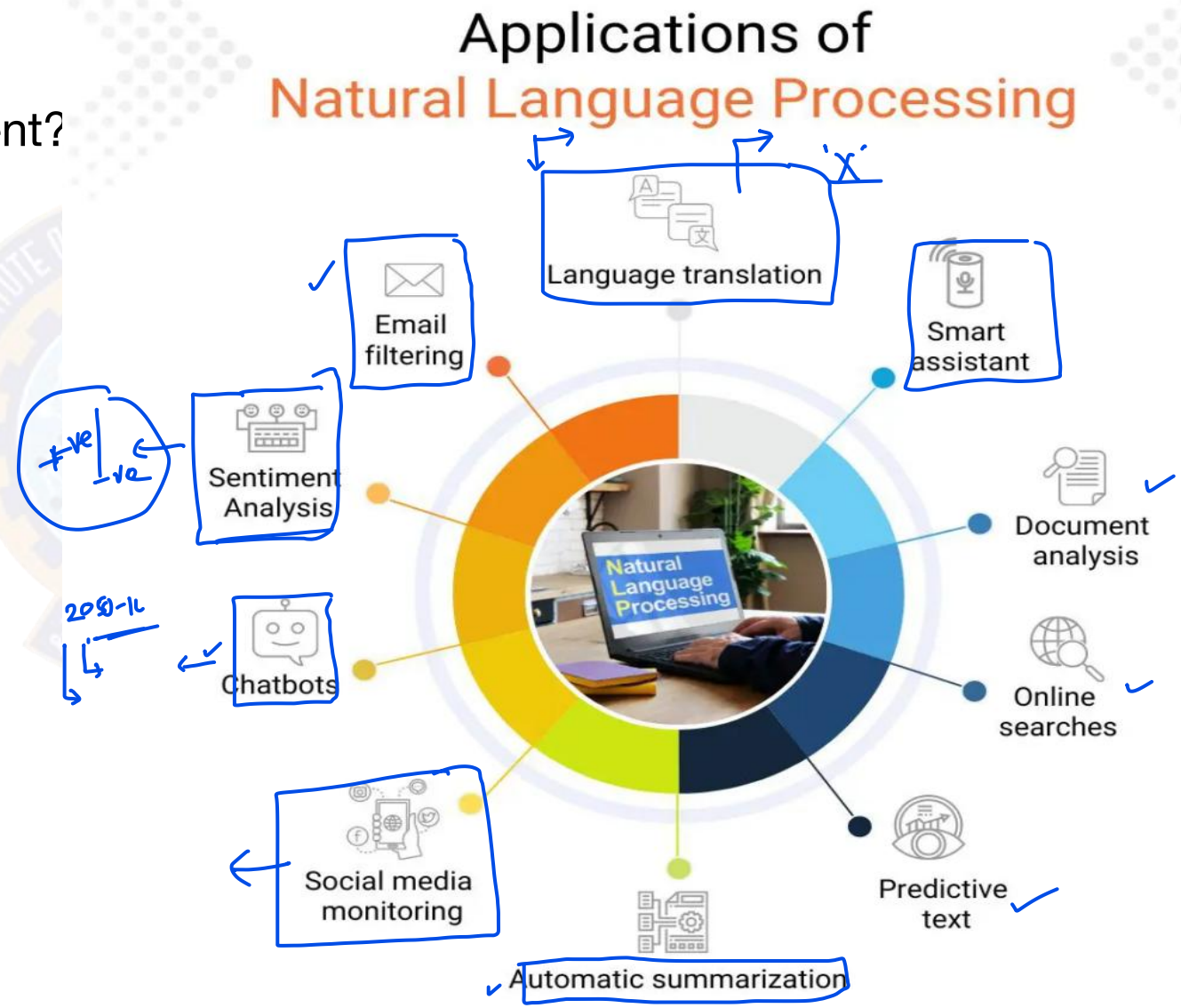
HAL:"I can tell from the tone of your voice, Dave, that you're upset."
HAL:"Why don't you take a stress pill and get some rest?"
HAL:"Can you hold it a bit closer?... That's Dr. Hunter, isn't it?"

To achieve this level of intelligence, an AI system needs

- Emotion Recognition – detecting emotions from speech and text.
- Multimodal Understanding – integrating language, speech, images, and vision.
- Context & Situational Awareness – resolving ambiguity using surrounding context.
- World Knowledge – understanding facts, events, and everyday concepts.
- Semantic Understanding – interpreting the meaning of language.
- Intent Recognition – identifying the speaker's goals and purpose.
- Reasoning & Common Sense – making logical inferences and informed decisions.

NLP Applications

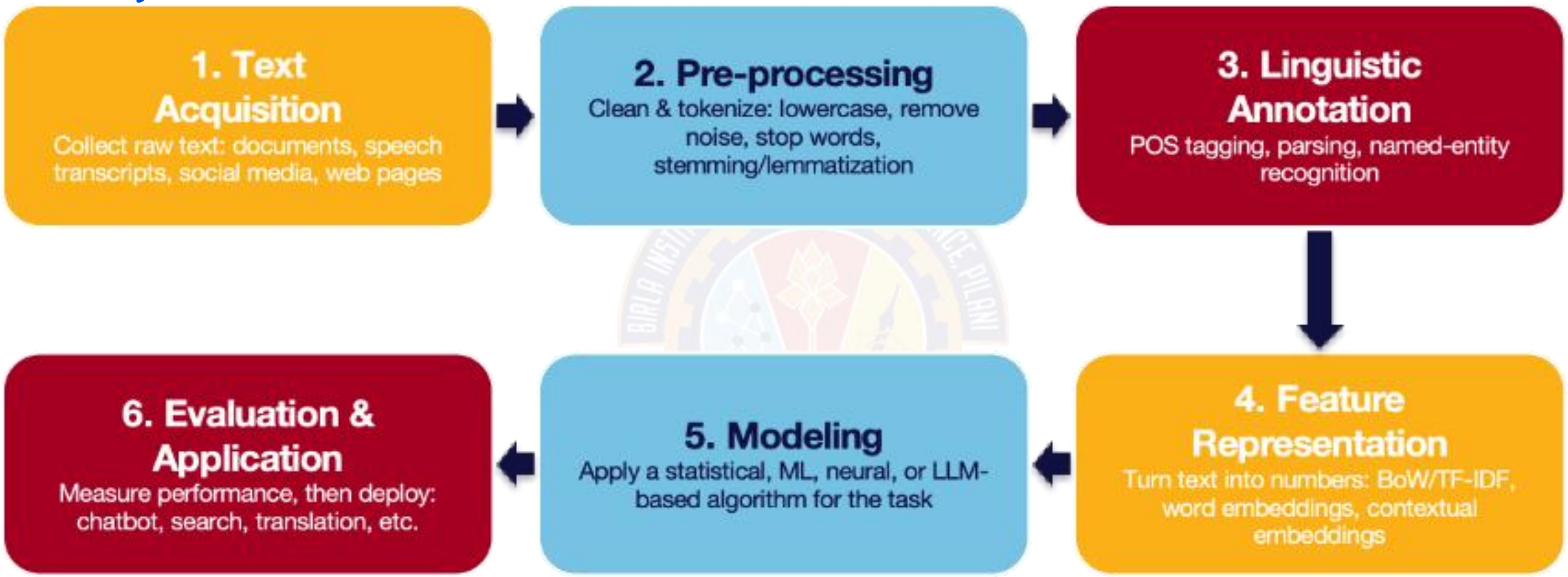
- Question answering
  - Who is the first Taiwanese president?
- Text Categorization/Routing
  - e.g., customer e-mails.
- Text Mining
  - Find everything that can be done with NLP
- Machine (Assisted) Translation
- Language Teaching/Learning
  - Usage checking
- Spelling correction
  - Is that just dictionary lookup?



Application areas

- Text-to-Speech & Speech recognition
- Healthcare
- Natural Language Dialogue Interfaces to Databases
- Information Retrieval_
- Information Extraction (http://nlp.stanford.edu:8080/ner/process)
- Document Classification
- Document Image Analysis
- Automatic Summarization (https://quillbot.com/summarize)
- Text Proof-reading – Spelling & Grammar_
- Machine Translation_
- **Fake News and Cyberbullying Detection**
- **Monitoring Social Media Using NLP**
- Plagiarism detection
- Look-ahead typing / Word prediction_
- Question Answering System (http://start.csail.mit.edu/index.php)
- Sentiment Analysis (https://komprehend.io/sentiment-analysis)

The NLP Pipeline


1. Text Acquisition
Collect raw text: documents, speech transcripts, social media, web pages
2. Pre-processing
Clean & tokenize: lowercase, remove noise, stop words, stemming/lemmatization
3. Linguistic Annotation
POS tagging, parsing, named-entity recognition
4. Feature Representation
Turn text into numbers: BoW/TF-IDF, word embeddings, contextual embeddings
5. Modeling
Apply a statistical, ML, neural, or LLM-based algorithm for the task
6. Evaluation & Application
Measure performance, then deploy: chatbot, search, translation, etc.

Note: since ~2020, large pretrained LLMs increasingly fold stages 2–5 into a single model call (prompting) — but understanding each stage is still essential to debug, evaluate, and trust an NLP system.

Pipeline in Action: A Worked Example

Stage	What Happens to Our Example: “The movie was not bad at all!”
1. Text Acquisition	Input review text is collected from a movie review website
2. Pre-processing	Lowercase + remove punctuation → tokens: [the, movie, was, not, bad, at, all]
3. Linguistic Annotation	POS tags: DT NN VBD RB JJ IN DT; negation scope detected around “not bad”
4. Feature Representation	Convert to embeddings / TF-IDF vector — captures that “not bad” ≠ “bad”
5. Modeling	A classifier (logistic regression, neural net, or LLM prompt) predicts sentiment
6. Evaluation & Application	Output: Positive sentiment (94% confidence) → shown on the review dashboard

Modern NLP & Generative AI Ecosystem

Foundation Models & AI Assistants

ChatGPT (OpenAI) – Conversational AI, content generation, coding, reasoning
Google Gemini – Multimodal AI for text, images, audio, and video
Claude (Anthropic) – Long-context reasoning and document understanding
GitHub Copilot – AI-powered coding assistant

Open-Source NLP & LLM Frameworks

Hugging Face Transformers – Pre-trained NLP and LLM models
spaCy – Industrial-strength NLP pipeline
NLTK – Educational NLP toolkit
Gensim – Topic modeling and document similarity
LangChain / LlamaIndex – Building LLM applications and RAG pipelines

Cloud AI Platforms

OpenAI API
Google Vertex AI / Gemini API
Amazon Bedrock
Azure AI Foundry
IBM watsonx

Supporting Technologies

- Vector Databases (FAISS, ChromaDB, Pinecone)
- Embedding Models
- Retrieval-Augmented Generation (RAG)
- Knowledge Graphs
- Speech-to-Text & Text-to-Speech APIs

Why is NLP Big Deal

- L = Words + rules + exceptions..
- Ambiguity at all levels..
- We speak different languages..
- And language is a cultural entity..
- So they are not equivalent..
- Highly systematic but also complex..
- Keeps changing.. New words, New rules and New exceptions..
- Source : Electronic texts / Printed texts / Acoustic Speech Signal.. they are noisy..
- Language looks obvious to us.. But it is a Big Deal ☺!

Why is NLP difficult?

Where is Black Panther playing in Mountain View?

Black Panther is playing at the Century 16 Theater.

When is it playing there?

It's playing at 2pm, 5pm, and 8pm.

OK. I'd like 1 adult and 2 children for the first show.
How much would that cost?

Need domain knowledge, discourse knowledge, world knowledge

financial org

I go to bank for deposit

I go to bank to relax

riverbank

↳ How →

↳ NLP → "Computer"

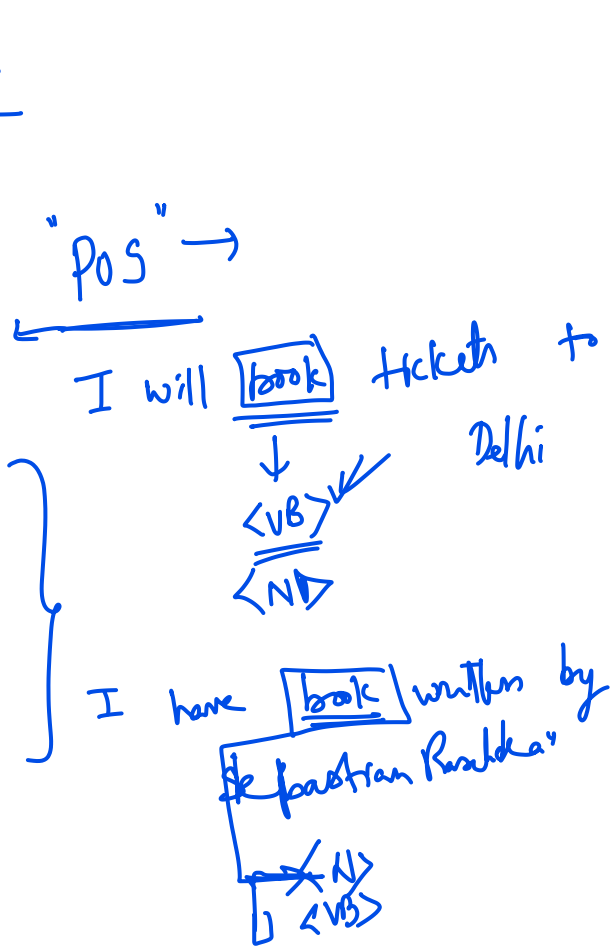
↳ Numbers [1|0's]

↳ Text

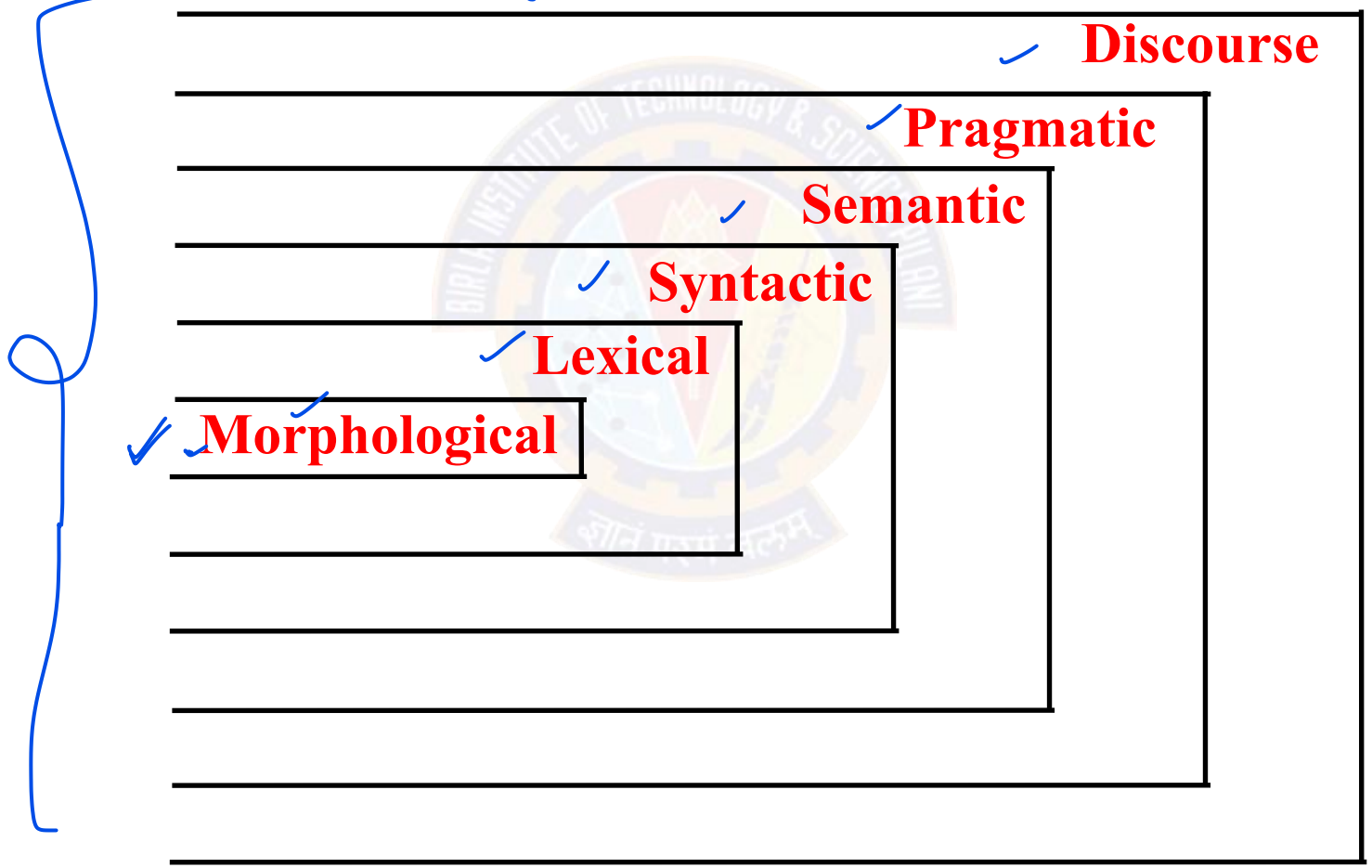
↳ "Embedding"

Types of Ambiguities

- **Structural Ambiguities**
  - Namrata thinks she understands me.
  - She thinks Namrata understands me.
  - Visiting relatives can be nuisance. (two meanings)
- **Grammatical Ambiguities**
  - I (feminine or masculine) go.
  - Can- Noun = container, Can – Modal(auxiliary verb),
  - Can-verb = to can means to pack etc
- **Lexical Ambiguities**
  - Polysemy Ex: "understand" (I get it)
  - Homonymy Ex: Bank= river, financial bank



Different Levels of Language Analysis



Levels of language understanding

- **Morphological Knowledge :** Concerns how words are constructed from basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language (Ex: "*friendly*" is derived from the meaning of noun "*friend*" and suffix "*-ly*", which transforms noun into adjective)

- **Lexical Knowledge:** Concerns with listing of words and categorizing them Ex: friendful or beautyship is incorrect lexically. But friendship and beautiful is correct

- **Syntactic Knowledge :**Concerns how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subpart of other phrases Ex: "Large have green ideas nose" is lexically correct but syntactically incorrect.

Levels of language understanding

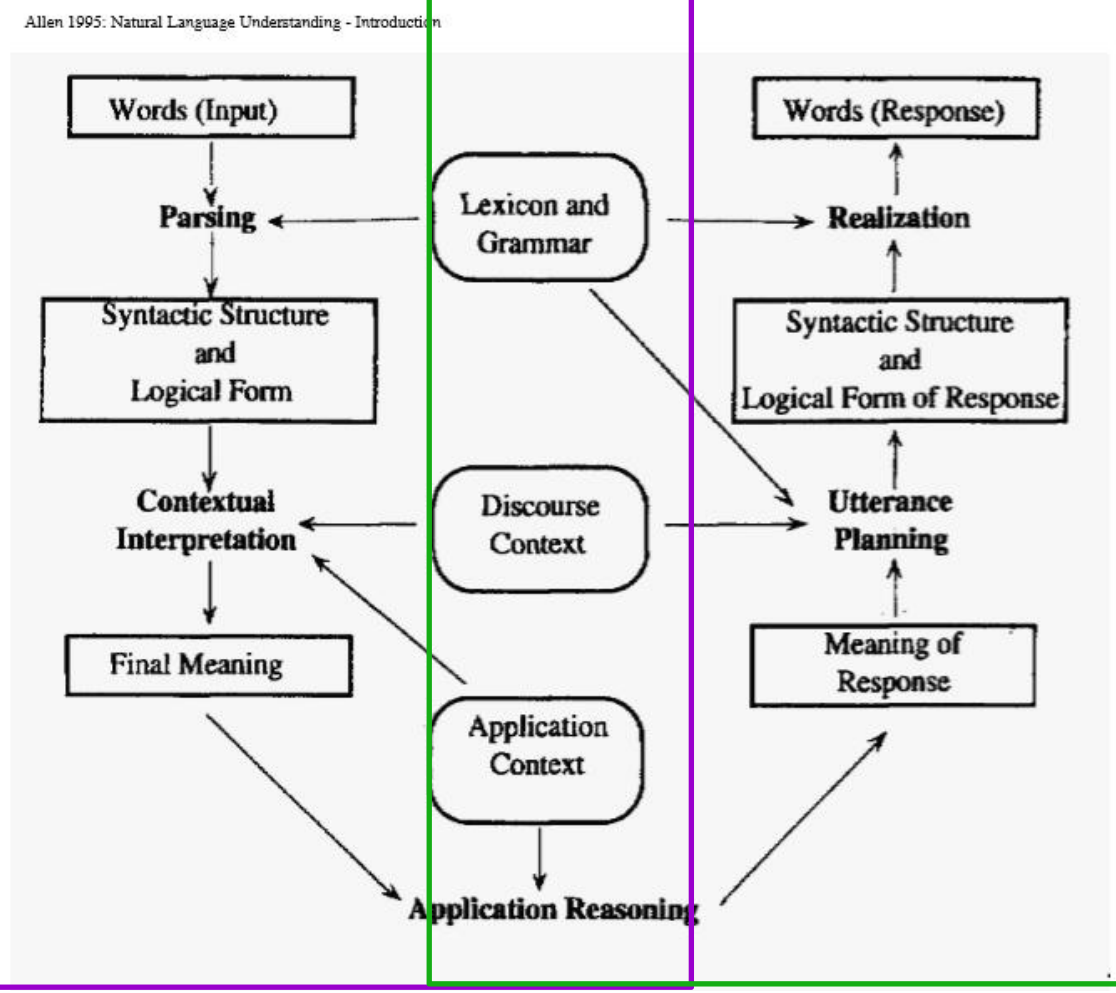
- **Semantic Knowledge :** Concerns what words mean and how these meanings – combine in sentences to form sentence meanings. This is the study of context-independent meaning.
  - Ex: “Green ideas have large noses” is syntactically correct but semantically incorrect.

- **Pragmatic Knowledge** :Concerns how sentences are used in different situations how use affects the interpretation of sentence
  - “She cuts banana with a pen” is semantically correct but pragmatically incorrect as it has no useful meaning.

- **Discourse Knowledge** :Concerns how the immediately preceding sentences affect the interpretation of next sentence
  - Ex: Chetana completed PhD student at IIT Bombay. She is a Professor at BITS Pilani.

The Organization of Natural Language Processing Systems

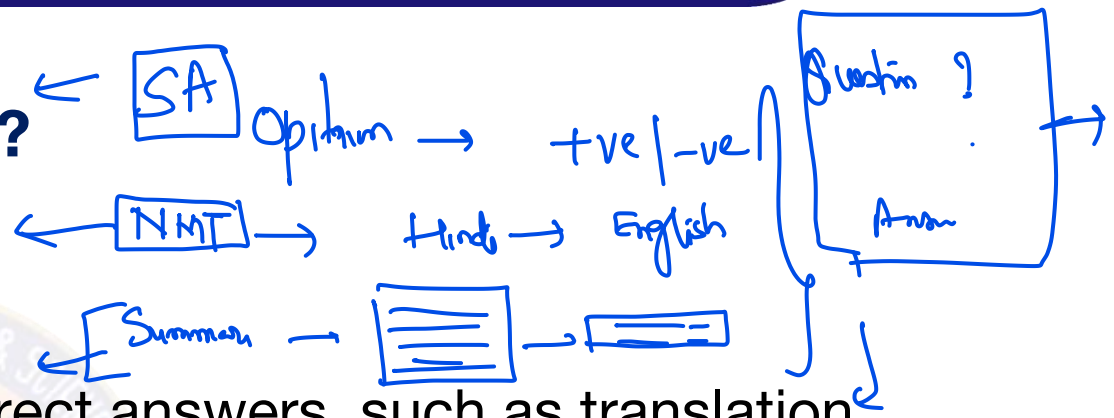
Natural Language Understanding



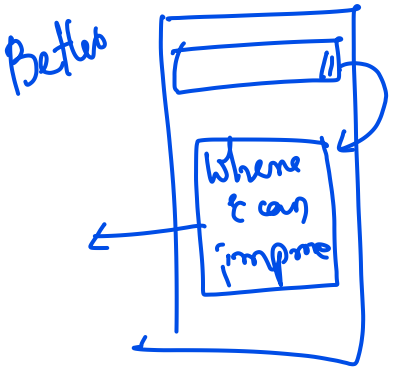
Natural Language Generation

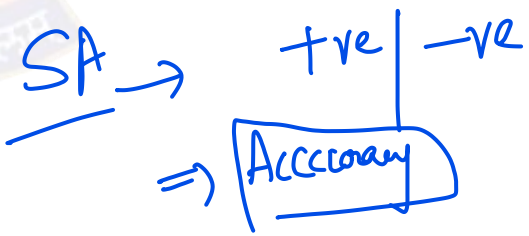
Evaluating Language Understanding Systems

How Do We Know an NLP System is Good?



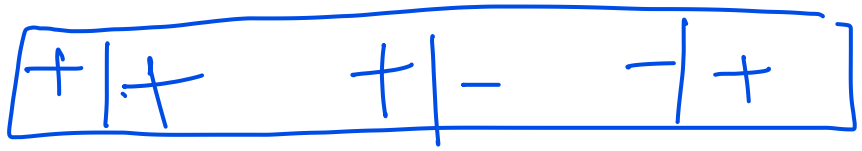
- What metrics to use?
- How do we evaluate tasks with multiple correct answers, such as translation or summarization?
- Are the human judgments measuring something real? reliable?
- Is our evaluation dataset representative of real-world language?
- How reliable or certain are the results?





"Evaluation"

Building an NLP system is only half the challenge. The real question is: Can we trust its output?



Contingency Table

	Actual: Positive	Actual: Negative
Predicted: Positive	True Positive (TP)	False Positive (FP)
Predicted: Negative	False Negative (FN)	True Negative (TN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Accuracy = (TP + TN) / (TP + FP + FN + TN)

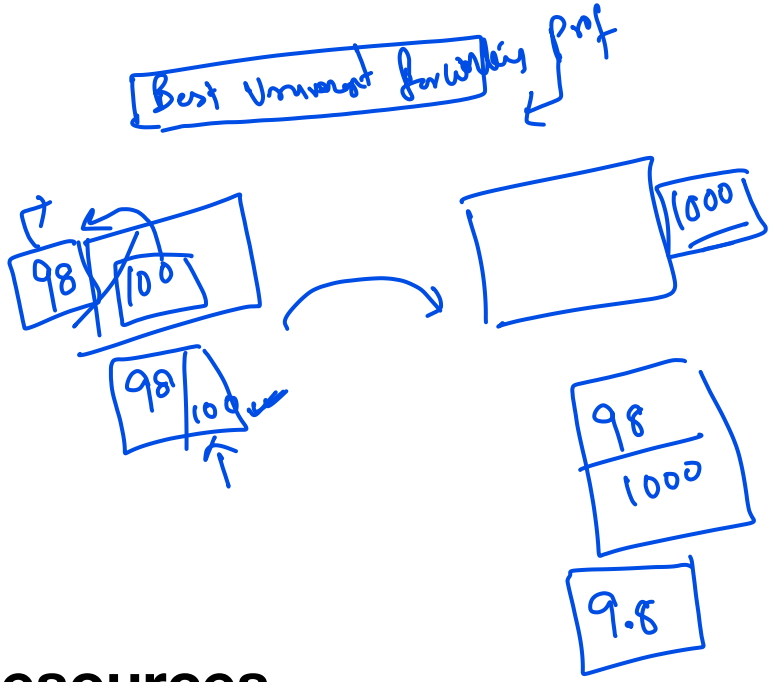
F1 Score = 2 × (Precision × Recall) / (Precision + Recall) — NEW: balances precision & recall in one number

Evaluating Generation: BLEU, ROUGE & Beyond

Metric	What It Measures	Where It's Used
Perplexity	How well a language model predicts held-out text (lower = better)	Language Modeling (N-gram, Neural LMs)
BLEU	N-gram precision overlap between generated and reference translations	Machine Translation
ROUGE	N-gram/sequence recall overlap between generated and reference summaries	Text Summarization
F1 Score	Harmonic mean of precision and recall (see Contingency Table)	Classification tasks (POS tagging, sentiment, etc.)
BERTScore (model-based)	Semantic similarity using contextual embeddings, not exact word overlap	Any generation task — catches paraphrases BLEU/ROUGE miss
LLM-as-Judge / Human Preference	Using another LLM or human raters to score or compare outputs	Modern LLM & RAG evaluation (chatbots, dialogue)

References

- Books
  - Jurafsky & Martin – Speech and Language Processing
  - Manning & Schütze – Foundations of Statistical NLP
  - Eisenstein – Natural Language Processing
- University Courses
  - Stanford CS224N
  - Stanford CS224U
  - Edinburgh NLP Course
- Documentation
  - Hugging Face Learn
  - OpenAI Platform Docs
  - spaCy Documentation
  - NLTK Book



Modern NLP Resources

- Hugging Face Course
- Hugging Face Transformers Documentation
- OpenAI API Documentation
- spaCy Documentation
- NLTK Book

Thank You!